DOCUMENT RESUME

ED 268 140 TM 850 783


AUTHOR Dorans, Neil J.


TITLE On Correlations, Distances and Error Rates.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-85-32
PUB DATE Jul 85
NOTE 34p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC0^ Plus Postage.
DESCRIPTORS *Classification; *Correlation; *Error of Measurement; *Estimation (Mathematics); Least Squares Statistics; *Prediction; Regression (Statistics); Statistical Studies; Validity
IDENTIFIERS *Error Analysis (Statistics); Mahalanobis Distance Function; *Shrunken Generalized Distance



ABSTRACT

The nature of the criterion (dependent) variable may play a useful role in structuring a list of classification/prediction problems. Such criteria are continuous in nature, binary dichotomous, or multichotomous. In this paper, discussion is limited to the continuous normally distributed criterion scenarios. For both cases, it is assumed that the predictor variables are continuous multivariate normal. For the binary variable case, the multivariate normal assumption is conditioned on the binary criterion, that is, for each value of the binary criterion, the predictors are multivariate normal with a common covariance matrix, but different centroids. In other words, for the continuous criterion case, the correlations model is used, while for the binary case the assumptions associated with the classic two-group discriminant analysis problem are employed. When these two models fit some population of data, then the use of standard loss functions yields well known population-optimal solutions. A unified framework for classification and prediction problems is presented. Well known and lesser known relationships among correlations, distances and error rates are established. A new population distance, the shrunken generalized distance, and a new estimator of the actual error rate are introduced. (Author/PN)

RESEARCH REPORT

# ON CORRELATIONS, DISTANCES AND ERROR RATES

Neil J. Dorans



Educational Testing Service
Princeton, New Jersey
July 1985

# ·relations, Distances and Error Rates[1]

Neil J. Dorans

College Board Statistical Analysis

July 1985

[1]The typing assistance of Debbie Giannacio is much appreciated, especially her work on the appendices.

## Abstract

A unified framework for classification and prediction problems is presented. Well known and lesser known relationships among correlations, distances and error rates are established. A new population distance, the shrunken generalized distance, and a new estimator of the actual error rate are introduced.

Classification and prediction problems abound. An extensive list of prediction and classification examples is easy to generate. Such a list could be structured by searching for similarities and identifying differences among the examples on it. Ultimately, each entry on the list could be viewed as a member of one of a smaller set of classes of prediction/classification problems.

The nature of the criterion (dependent) variable may play a useful role in structuring a list of classification/prediction problems. Some criteria are essentially continuous in nature, e.g., scores on a long test. Other criteria are inary, e.g., group membership. Other criteria appear binary but may be thought of as dichotomizations of a continuous underlying criterion, e.g., pass/fail grading of an essay. Other criteria are multichotomous. In this paper, discussion is limited to the continuous normally distributed criterion sceranios. For both cases, it will be assumed that the predictor variables are continuous multivariate normal. For the binary variable case, the multivariate normal assumption is conditioned on the binary criterion, i.e., for each value of the binary criterion, the predictors are multivariate normal with a common covariance matrix, but different centroids. In other words, for the continuous criterion case, the correlations model is used, while for the binary case the assumptions associated with the classic two-group discriminant analysis problem are employed.

When these two models fit some population of data, then the use of standard loss functions (least squares in the continuous criterion case; maximum probability in the binary criterion case) yields well known population-optimal solutions.

## Population Indices

### The Continuous Criterion Case

For the continuous case, ordinary least squares regression yields the population optimal regression equation,

(1) $\underline{\beta}_p = \Sigma_{xx}^{-1}\underline{\sigma}_{xy}$ ,

where $\Sigma_{xx}$ is the r-by-r population covariance matrix among the predictors (X), and $\underline{\sigma}_{xy}$ is the r-by-1 vector of covariances between the predictors and criterion y. The population multiple correlation, $\rho_p$, or validity coefficient,

(2) $\rho_p = (\underline{\beta}_p'\underline{\sigma}_{xy})/(\underline{\beta}_p'\Sigma_{xx}\underline{\beta}_p\sigma_y^2)^{1/2}$ ,

indexes the extent to which the predicted criterion orderings,

(3) $\hat{y}_p(\underline{x}_i) = \underline{\beta}_p'\underline{x}_i + \beta_o$ ,

obtained by applying the regression weights to the r predictor scores for the individual ($\underline{x}_i$), matches the ordering of the criterion scores in the population. And, the population mean squared error, $MSE_p$, indexes how accurately the predicted criterion scores match the actual criterion scores in the population,

(4) $MSE_p = \varepsilon(y_p - \hat{y}_p)^2$ .

The population squared validity and mean squared error of prediction are related via

(5) $\rho_p^2 = 1 - MSE_p/\sigma_y^2$ .

### The Binary Criterion Case

In the standard two subpopulation classification case in which the subpopulations, $K_1$ and $K_2$, are of equal size, i.e., $pr(K_1) = pr(K_2) = .5$,

and the n predictors in both subpopulations follow a multivariate normal distribution with the same covariance matrix, $\Sigma$, and different centroids, $\underline{\mu}_1$ and $\underline{\mu}_2$, optimal classification according to the maximum probability, maximum likelihood, and generalized distance rules (Huberty, 1975; Tatsuoka, 1971) all reduce to assignment to the subpopulation with the nearest centroid. Operationally, this is accomplished by computing the Wald-Anderson classification statistic

$$(6) \qquad W_p(\underline{x}_i) = \underline{\lambda}_p' [\underline{x}_i - .5(\underline{\mu}_1 - \underline{\mu}_2)] ,$$

where $\underline{\lambda}_p$ is r-by-1 vector containing Fisher's (1936) population linear discriminant weights,

$$(7) \qquad \underline{\lambda}_p = \Sigma^{-1}(\underline{\mu}_1 - \underline{\mu}_2) .$$

The adequacy of classification in the population is indexed by the optimal error rate (Hills, 1966),

$$(8) \qquad E_p = .5\ \Phi\left[\frac{-W_p(\underline{\mu}_1)}{(\underline{\lambda}_p'\Sigma\,\underline{\lambda}_p)^{1/2}}\right] + .5\ \Phi\left[\frac{W_p(\underline{\mu}_2)}{(\underline{\lambda}_p'\Sigma\,\underline{\lambda}_p)^{1/2}}\right] ,$$

which is the probability of misclassification associated with use of the population optimal classification rule, $W_p$. In (8), $\Phi$ is the standard normal distribution function. It has been shown that $E_p$ can be expressed in terms of the separation between $K_1$ and $K_2$, the population Mahalonobis (1936) or generalized distance,

$$(9) \qquad \delta_p^2 = (\underline{\mu}_1 - \underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1 - \underline{\mu}_2) ,$$

which can be thought of as the squared standardized difference between populations $K_1$ and $K_2$ along the dimension defined by $\underline{\lambda}_p$,

$$(10) \qquad \delta_p^2 = [\underline{\lambda}_p'(\underline{\mu}_1 - \underline{\mu}_2)^2/(\underline{\lambda}_p'\Sigma\underline{\lambda}_p)].$$

In particular, $E_p$ can be expressed as a function of $\delta_p^2$

(11) $$E_p = .5\Phi\left[\frac{-.5\,\delta_p^2}{\delta_p}\right] + .5\Phi\left[\frac{-.5\,\delta_p^2}{\delta_p}\right] = \Phi\,[-.5\delta_p] \ ,$$

which can be obtained by evaluating (6) at $\underline{\mu}_1$ and $\underline{\mu}_2$ in (8), and simplying the expression using (9) and (10).

Parallels Between Continuous Criterion and Binary Criterion Cases

There are parallels between the continuous criterion case and the binary criterion case. There are parallel sets of weights: $\underline{\beta}_p$ for the continuous criterion case, $\underline{\lambda}_p$ for the binary criterion case. The squared correlation measure of association parallels the generalized distance $\delta_p^2$. And the mean squared error of prediction, $MSE_p$, parallels the optimal error rate, $E_p$. In fact, for the binary criterion case, $\underline{\beta}_p$ and $\underline{\lambda}_p$ are known to be proportional. In addition, it can be shown that $\delta_p^2$ and $\rho$ are related (See Appendix A),

(12) $$\delta_p^2 = \left[\frac{\rho^2}{1-\rho^2}\right] / \left[pr(K_1) \cdot pr(K_2)\right] \ .$$

Cross-Validity and the Actual Error Rate

In practice, we seldom work with populations. Instead, we are limited to samples of data. Substitution of sample mean, variances and covariances into (1) - (12) produces sample analogues of $\beta_p$, $\underline{\lambda}_p$, $\rho_p$, $\delta_p^2$, $MSE_p$ and $E_p$. For example, for the continuous criterion case, we have

(13) $$\underline{b}_s = C_{xx}^{-1}\,\underline{c}_{xy} \ ,$$

where $C_{xx}$ and $\underline{c}_{xy}$ are the sample analogues of $\Sigma_{xx}$ and $\underline{\sigma}_{xy}$. For the binary criterion case, we have

(14) $\quad \underline{b}_s = G^{-1}(\bar{\underline{x}}_1 - \bar{\underline{x}}_2)$ ,

where G, $\bar{\underline{x}}_1$ and $\bar{\underline{x}}_2$ are the sample within-group covariance matrix and sample centroids, respectively. In general the usefulness of a regression equation or a classification rule should be assessed by its performance in the population, not its performance in the sample. For the continuous criterion case, the population cross-validity coefficient,

(15) $\quad R_c = (\underline{b}_s' \underline{\sigma}_{xy})/(\underline{b}_s' \Sigma_{xx} \underline{b}_s \sigma_y^2)^{1/2}$ ,

and the mean squared error of prediction $MSE_c$ associated with use of the sample weights, $\underline{b}_s$, in the population index the long-term usefulness of the sample weights. Lord (1950) developed an estimator for the $MSE_c$ for when the predictors are considered fixed, i.e., the regression model, while Stein (1960) developed an estimator for the $MSE_c$ under the correlation model. Browne (1975), as demonstrated by Drasgow, Dorans and Tucker (1979) and Drasgow and Dorans (1982), developed an estimator of the population squared cross-validity that is virtually unbiased and robust to violations of multivariate normality in the predictors.

For the binary criterion case, the actual error rate, $E_c$, summarizes how well a sample classification rule works in the population. The actual error rate is the probability of misclassification associated with use of the sample classification rule in the population. In many ways, the actual error rate is more important than the optimal error rate. The actual error rate is akin to the population mean squared error rate associated with a sample regression equation. In the two equal-sized subpopulation case under consideration, the expression for the actual error rate is

$$(16) \qquad E_c = .5\Phi\left[\frac{-W_s(\underline{\mu}_1)}{V_w}\right] + .5\Phi\left[\frac{W_s(\underline{\mu}_2)}{V_w}\right] ,$$

where $W_s(\underline{\mu}_k)$ is the sample classification statistic $W_s(\underline{x}_i)$ evaluated at $\underline{\mu}_k$,

$$(17) \qquad W_s(\underline{x}_i) = \underline{l}_s'[\underline{x}_i - .5(\bar{\underline{x}}_1 - \bar{\underline{x}}_2)] ,$$

and $V_W$ is the variance in each subpopulation of the composite defined by the sample discriminant weights, $\underline{l}_s$,

$$(18) \qquad V_W = \underline{l}_s'\Sigma\,\underline{l}_s .$$

The literature contains several estimators of the actual error rate for the two multivariate normal subpopulation case. One class of estimators, that are somewhat heuristic, are the distance-modification estimators. This class of estimators attempt to mimic the relationship between $E_p$ and $\delta_p^2$ stated in (11) by substituting distance estimates into

$$(19) \qquad \hat{E}_c = \Phi[(-.5\ D)].$$

Two of the most popular distance modification methods are the D-method and the DS-method. The D-method uses the sample generalized distance $D_s^2$ for $D^2$ in (19). The DS-method uses

$$(20) \qquad D_{DS}^2 = (N-n-3)D_s^2/(N-2) ,$$

which is the positive portion of an unbiased estimate of $\delta_p^2$,

$$(21) \qquad \hat{\delta}_p^2 = D_{DS}^2 - Nn/(N/2)^2 .$$

According to Lachenbruch and Mickey (1968), $D_{DS}^2$ is used instead of $\hat{\delta}_p^2$ to avoid negative distance estimates.

The Shrunken Generalized Distance

While attracted by the intuitive appeal of these two distance-modification procedures, I am convinced that they are inappropriate, i.e., not the right distances. $D_s^2$ is like the sample squared multiple correlation, $R_s^2$; in fact they can be related. Using a positively biased estimate of the population Mahalonbis distance, as the DS-method does, is like using a positively biased estimate of the population squared multiple correlation $\rho_p^2$ to estimate the population squared cross-validity coefficient $R_c^2$. An estimate of some distance that was analogous to the squared cross-validity is clearly needed. So I invented (Dorans, 1979) the shrunken generalized distance, $D_c^2$, between two subpopulation centroids, $\underline{\mu}_1$ and $\underline{\mu}_2$.

The shrunken generalized distance is the squared standardized distance between the projections of the two subpopulation centroids onto the dimension defined by the sample discriminant weights, $\underline{l}_s$. These projections are obtained by evaluating the sample classification statistic in (17) at $\underline{\mu}_1$ and $\underline{\mu}_2$. The variance along this dimension is that defined in (18). The shrunken generalized distance is formally expressed as

$$(22) \qquad D_c^2 = (W_s(\underline{\mu}_1) - W_s(\underline{\mu}_2))^2 / V_W \ ,$$

which can be rewritten as

$$(23) \qquad D_c^2 = \underline{l}_s' (\underline{\mu}_1 - \underline{\mu}_2)^2 / (\underline{l}_s' \Sigma \underline{l}_s) \ .$$

To appreciate what $D_c^2$ represents, it is helpful to resort to geometric imagery. For the case of two multivariate normal subpopulations with equal covariance matrices and different centroids, the population discriminant weights $\underline{\lambda}_p$ define the dimension in the n-dimensional predictor space along

which there is maximal separation between the subpopulations. As noted earlier, the population generalized distance, $\delta_p^2$, can be thought of as the squared standardized distance between the subpopulation centroids' projections on this dimension defined by $\underline{\lambda}_p$ (See equation 10).

Suppose that instead of $\underline{\lambda}_p$, we had used the sample weights $\underline{l}_s$ to define a dimension in the population. When the centroids of the two subpopulations are projected onto this dimension, two means are produced, one for each subpopulation on this dimension. The squared standardized difference between these means is the shrunken generalized distance. Unless the dimension defined by the sample weights is parallel or collinear to the dimension defined by the population optimal weights, this squared standardized difference in means will be smaller than the population generalized distance. In other words, the distance will have shrunken; hence, the phrase shrunken generalized distance.

This shrunken generalized distance should estimate the actual error rate better the modified distances used by the D-method and the DS-method. An estimator of the shrunken generalized distance was derived (See Appendix B),

$$(24) \qquad \hat{\Delta}_c^2 = \frac{\delta_p^2 \quad N-3 + N_1N_2N^{-1}(N-r-2)\ \delta_p^2}{(N-3)\ (r + N_1N_2N^{-1}\delta_p^2)} \quad ,$$

which uses the unbiased estimator of the population generalized distance defined in (21) and where $N_1$ and $N_2$ are the sample sizes for each subgroup, i.e., $N = N_1 + N_2$. A simulation was conducted to compare this new shrunken distance estimator with the two other distance modification estimators, as

well as five other estimators. I expected the MU-estimator, as I called it, to be superior to the D-method and DS-method because it used the appropriate distance, the shrunken generalized distance.

One of the five other estimators examined in the simulation is the OS-method, which is based on Okamoto's (1963) asymptotic expansion of the distribution of the sample Wald-Anderson statistic, $W_s$. Previous research (Lachenbruch and Mickey, 1968; Sorum, 1972) had demonstrated that the OS-method was the best estimator available. The equal N special case of Okamoto's OS-method was used,

$$E_c(OS) = \Phi(-.5D_{DS}) + \Phi(.5D_{DS})\left[\frac{(r-1)}{ND_{DS}} + \frac{D_{DS}}{4N} + \frac{(r-1)D_{DS}}{4(N-2)}\right]. \quad (25)$$

In (25), $\Phi$ is the standard normal density function.

The simulation study (Dorans, 1984) demonstrated that the MU-method is the best of the heuristic distance - modification procedures. In addition, it seemed to perform as well as if not better than the OS-method. The MU-method works well because it is an estimator of the minimum actual error rate associated with use of the sample classification rule in the population. (See Appendix C for proof of this statement.)

## The Shrunken Generalized Distance and the Squared Cross-Validity

In Appendix A, it is demonstrated that the population parameters $\rho^2$ and $\delta_p^2$ are related as in equation (12). In Appendix D. the relationship between the shrunken generalized distance, $D_c^2$, and the squared cross-

validity coefficient, $R_c^2$, is shown to be

(26) $$D_c^2 = \frac{R_c^2}{(1-R_c^2)} \Big/ (q_1 q_2) ,$$

where $q_1$ and $q_2$ are the relative sizes of subpopulations $K_1$ and $K_2$. This relationship between $R_c^2$ and $D_c^2$ completes the unified framework for classification and prediction problems.

The framework distinguishes between continuous criterion cases and truly binary criterion cases. On the continuous side of the ledger we have $\underline{\beta}_p$, $\underline{\rho}_p$ and $MSE_p$ with (2), (3) and (5) serving as definitions and establishing relationships. On the binary side we have the analogous $\underline{\lambda}_p$, $\delta_p^2$, and $E_p$ with (7), (8), (9) and (10) serving as defining relationships. Then Appendix A demonstrates that $\delta_p^2$ and $\rho_p^2$ are related as in (12).

The framework includes the use of sample weights in the population. For the continuous criterion case, we have $R_c^2$ and $MSE_c$. For the binary criterion case, we have $D_c^2$ and $E_c$. Appendix D establishes the relationship between $R_c^2$ and $D_c^2$, while Appendix C shows how $D_c^2$ and $E_c$ may be related.

In order to complete the framework for prediction/classification problems, the notion of the shrunken generalized distance, $D_c^2$, was introduced. In addition to being the missing piece in the analytic framework, this distance is useful for estimating the actual error rate, as demonstrated elsewhere (Dorans, 1984).

## References

Anderson, T.W. (1958). An introduction to multivariate statistical analysis. New York: John Wiley & Sons, Inc.

Browne, M. W. (1975). Predictive validity of a linear regression equation. British Journal of Mathematical and Statistical Psychology, 28, 79-87.

Dorans, N. J. (1979). Reduced rank classification and estimation of the actual error rate (Doctoral Dissertation, University of Illinois, 1978). Dissertation Abstracts International, 39 (12), 6095-B. (University Microfilms Order No. 79-13434).

Dorans, N. J. (1984). The shrunken generalized distance: A useful concept for estimation of the actual error rate. (RR-84-1). Princeton, NJ: Education Testing Service.

Drasgow, F., Dorans, N.J., & Tucker, L.R. (1979). Estimators of the squared cross-validity: A monte carlo investigation. Applied Psychological Measurement, 3, 387-399.

Drasgow, F. & Dorans, N.J. (1982). Robustness of estimators of the squared multiple correlation and squared cross-validity coefficient to violations of multivariate normality. Applied Psychological Measurement, 6, 185-200.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Annuals of Eugenics, 7, 179-188.

Hills, M. (1966). Allocation rules and their error rates. Journal of the Royal Statistical Society, Series B, 28, 1-31.

Huberty, C. J. (1975). Discriminant analysis. Review of Educational Research, 45, 543-598.

Kshirsagar, A. M. (1972). Multivariate analysis. New York: Marcel Dekker, Inc.

Lachenbruch, P. A. (1968). On expected values of probabilities of misclassification in discriminant analyses, necessary sample size, and a relation with the multiple correlation coefficient. Biometrics, 24, 823-834.

Lachenbruch, P. A. & Mickey, M. R. (1968). Estimation of error rates in discriminant analysis. Technometrics, 10, 1-11.

Lord, F. M. (1950). Efficiency of prediction when a regression equation from one sample is used in a new sample (RB-50-40). Princeton, NJ: Educational Testing Service.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. Proceedings of the National Institute of Science, India, 12, 49-55.

Okamoto, M. (1963). An asumptotic expansion for the distribution of the linear discriminant function. Annals of Mathematical Statistics, 34, 1286-1301. Correction: Annals of Mathematical Statistics, 1968, 39, 1358-1359.

Sorum, M. (1972). Estimating the expected and the optimal probabilities of misclassification. Techometrics, 14 (4), 935-943.

Stein, C. (1960). Multiple regression. In I. Olkin et. al. (Eds.), Contributions to probability and statistics. Stanford, CA: Stanford University Press.

Tatsoka, M. M. (1971). Multivariate analysis: Techniques for educational and psychological research. New York: Wiley.

Welch, B. L. (1939). Note on discriminant functions. Biometrika, 31, 218-220.

## APPENDIX A

### RELATIONSHIP BETWEEN $\rho_p^2$ AND $\delta_p^2$

In general, let $\Gamma$ be a (r+1)-by-(r+1) covariance having the form:

$$\text{(A.1)} \qquad \Gamma = \left[\begin{array}{c|c} \Sigma_{xx} & \underline{\sigma}_{xy} \\ \hline \underline{\sigma}_{yx} & \sigma_y^2 \end{array}\right]$$

where, $\underline{\sigma}_{yx}$ is a 1-by-r vector of covariances for the criterion variable, Y, with each of the r predictor variables X, $\Sigma_{xx}$ is the intercovariance matrix among the r predictors, and $\sigma_y^2$ is the variance of the criterion. When Y is a binary variable representing group membership, taking the value 1 if an individual is from subpopulation $K_1$, and the value 0 if an individual is from subpopulation $K_2$, $\sigma_y^2$ and $\underline{\sigma}_{yx}$ take on special forms. In particular, $\sigma_y^2$ is defined as the product

$$\text{(A.2)} \qquad \sigma_y^2 = q_1 q_2 \quad ,$$

where $q_1$ and $q_2$ are the proportions of individuals in $K_1$ and $K_2$ respectively. The covariance vector takes on the form

$$\begin{aligned} \text{(A.3)} \qquad \underline{\sigma}_{xy} &= q_1(1)(\underline{\mu}_1 - \underline{\mu}) + q_2(0)(\underline{\mu}_2 - \underline{\mu}) \\ &= q_1(\underline{\mu}_1 - q_1\underline{\mu}_1 - q_2\underline{\mu}_2) \\ &= q_1 q_2(\underline{\mu}_1 - \underline{\mu}_2) \end{aligned}$$

where $\underline{\mu}_1$ is the r-by-1 centroid vector in $K_1$ while $\underline{\mu}_2$ is the r-by-1 centroid vector in $K_2$ and $\underline{\mu}$ is the grand mean

$$\text{(A.4)} \qquad \underline{\mu} = q_1\underline{\mu}_1 + q_2\underline{\mu}_2 \quad .$$

Therefore, when group membership is coded as a binary variable, the general expression for $\Gamma$ in (A.1) has the form

$$\text{(A.5)} \qquad \Gamma = \left[\begin{array}{c|c} \Sigma_{xx} & q_1 q_2(\underline{\mu}_1 - \underline{\mu}_2) \\ \hline q_1 q_2(\underline{\mu}_1 - \underline{\mu}_2)' & q_1 q_2 \end{array}\right]$$

In general, the population least squares regression weights $\underline{\beta}_p$ are defined as

$$(A.6) \qquad \underline{\beta}_p = \Sigma_{xx}^{-1}\underline{\sigma}_{xy} \quad ,$$

which, for a binary criterion variable reduces to

$$(A.7) \qquad \underline{\beta}_p = q_1 q_2 \, \Sigma_{xx}^{-1}(\underline{\mu}_1 - \underline{\mu}_2) \quad .$$

The population squared multiple correlation is defined as

$$(A.8) \qquad \rho^2 = (\underline{\beta}_p'\underline{\sigma}_{xy})^2/(\sigma_y^2 \underline{\beta}_p'\Sigma_{xx}\underline{\beta}_p)$$

$$= \frac{(q_1 q_2(\underline{\mu}_1 - \underline{\mu}_2)'\Sigma_{xx}^{-1}(\underline{\mu}_1 - \underline{\mu}_2)q_1 q_2)^2}{q_1 q_2(q_1 q_2(\underline{\mu}_1 - \underline{\mu}_2)'\Sigma_{xx}^{-1}\Sigma_{xx}\Sigma_{xx}^{-1}(\underline{\mu}_1 - \underline{\mu}_2)q_1 q_2)}$$

$$= q_1 q_2(\underline{\mu}_1 - \underline{\mu}_2)'\Sigma_{xx}^{-1}(\underline{\mu}_1 - \underline{\mu}_2)$$

The total covariance matrix among the predictors can be broken up into a within-groups covariance matrix $\Sigma$ and a between-groups covariance matrix,

$$(A.9) \qquad \Sigma_{xx} = \Sigma + q_1 q_2(\underline{\mu}_1 - \underline{\mu}_2)(\underline{\mu}_1 - \underline{\mu}_2)'$$

Substituting (A.9) into (A.8) yields

$$(A.10) \qquad \rho^2 = q_1 q_2(\underline{\mu}_1 - \underline{\mu}_2)' \, [\Sigma + q_1 q_2(\underline{\mu}_1 - \underline{\mu}_2)(\underline{\mu}_1 - \underline{\mu}_2)']^{-1}(\underline{\mu}_1 - \underline{\mu}_2)$$

$$= q_1 q_2(\underline{\mu}_1 - \underline{\mu}_2)'[\Sigma(I + q_1 q_2 \Sigma^{-1}(\underline{\mu}_1 - \underline{\mu}_2)(\underline{\mu}_1 - \underline{\mu}_2)')]^{-1}(\underline{\mu}_1 - \underline{\mu}_2)$$

$$= q_1 q_2(\underline{\mu}_1 - \underline{\mu}_2)' \, [I + q_1 q_2 \Sigma^{-1}(\underline{\mu}_1 - \underline{\mu}_2)(\underline{\mu}_1 - \underline{\mu}_2)']^{-1}\Sigma^{-1}(\underline{\mu}_1 - \underline{\mu}_2) \quad .$$

At this point, it is necessary to use the known matrix algebra relation (Kshirsagar, 1972),

$$(A.11) \qquad (I + LM)^{-1} = I - L(1 + ML)^{-1} M$$

Let,

$$(A.12) \qquad L = q_1 q_2 \Sigma^{-1}(\underline{\mu}_1 - \underline{\mu}_2)$$

and

$$(A.13) \qquad M = (\underline{\mu}_1 - \underline{\mu}_2)' \quad .$$

Then the relationship in (A.11) enables us to rewrite (A.10) as

$$(A.14)\quad \rho_2 = q_1q_2(\underline{\mu}_1-\underline{\mu}_2)'\left[I - \frac{q_1q_2\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2)(\underline{\mu}_1-\underline{\mu}_2)'}{1 + q_1q_2(\underline{\mu}_1-\underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2)}\right]\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2)$$

$$= q_1q_2(\underline{\mu}_1-\underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2) - \left[\frac{(q_1q_2)^2(\underline{\mu}_1-\underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2)(\underline{\mu}_1-\underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1-\underline{\mu}_2)}{1 + q_1q_2(\underline{\mu}_1 - \underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1 - \underline{\mu}_2)}\right].$$

By definition, the Manalanobis generalized distance between $\underline{\mu}_1$ and $\underline{\mu}_2$ equals

$$(A.15)\quad \delta_p^2 = (\underline{\mu}_1 - \underline{\mu}_2)'\Sigma^{-1}(\underline{\mu}_1 - \underline{\mu}_2) \quad .$$

When (A.15) is substituted into (A.14), the result is

$$(A.16)\quad \rho^2 = q_1q_2\delta_p^2 - (q_1q_2\delta_p^2)^2/(1 + q_1q_2\delta_p^2)$$
$$= (q_1q_2\delta_p^2 + (q_1q_2\delta_p^2)^2 - (q_1q_2\delta_p^2)^2)/(1 + q_1q_2\delta_p^2)$$
$$= (q_1q_2\delta_p^2)/(1 + q_1q_2\delta_p^2) \quad .$$

Thus,

$$(A.17)\quad \rho^2(1 + q_1q_2\delta_p^2) = q_1q_2\delta_p^2$$

and by simple rearrangement of terms, one obtains

$$(A.18)\quad \delta_p^2 = \frac{\rho^2}{1-\rho^2} \Big/ (q_1q_2) \quad .$$

# APPENDIX B

## AN ESTIMATOR OF THE SHRUNKEN GENERALIZED DISTANCE

### Minimum Actual Error Rate

The minimum actual error rate associated with a sample classification rule is the minimum probability of misclassification in the population associated with discrimination along the dimension defined by the sample discriminant weights. The minimum actual error rate is the minimum possible error rate associated with using a sample classification rule in the population and can be expressed as a function of the "shrunken" Mahalanobis distance,

$$\text{(B.1)} \qquad \min(E_c) = q_1\Phi\left[\frac{\ln(q_2/q_1) - .5D_c^2}{D_c}\right] + q_2\Phi\left[\frac{\ln(q_1/q_2) - .5D_c^2}{D_c}\right] .$$

(The derivation of (B.1), which assumes equal costs of misclassification, appears in Appendix C.) In the balance of this appendix, an estimator of $E(D_c^2)$ is developed. When substituted into (B.1) for $D_c^2$, this estimator can be used to approximate $\min(E_c)$, which in turn serves as an approximation to $E_c$, the actual error rate.

### An Estimator of $E(D_c^2)$ in Terms of $\delta_p^2$

If $D_c^2$ and the "relative sizes" of the two subpopulations, $q_1$ and $q_2$, are known, (B.1) could be used as a lower bound for the actual error rate associated with the use of a sample classification rule in the population. Unfortunately, $D_c^2$ is expressed in terms of the population parameters $\underline{\mu}_1$, $\underline{\mu}_2$, and $\Sigma$, and these quantities usually are unknown. (If these parameters

were known, concern about the actual error rate would be unnecessary since the optimal classification rule would be knowable.) An estimator for $D_c^2$ in terms of sample quantities is needed. In this section, the transformational invaria . $\delta_p^2$, the distribution for sample discriminant weights $\underline{l}_s$ and a logic paralleling that used by Drasgow, Dorans and Tucker (1979) are combined to obtain an estimator of $E(D_c^2)$ which can be used to estimate $D_c^2$.

Since the population generalized distance $\delta_p^2$ is invariant with respect to nonsingular transformations (Lachenbruch, 1975), it is possible to transform any r-dimensional space into an orthogonal orientation, in which the first dimension is parallel to the line passing through the subpopulation centroids and where each dimension has unit variance, without affecting $\delta_p^2$. In addition, $\delta_p^2$ is invariant to translations. These permissible transformations and translations can be applied to any two multivariate normal subpopulations with a common $\Sigma$ and centroids $\underline{\mu}_1$, and $\underline{\mu}_2$, to obtain a new covariance matrix $\tilde{\Sigma}$ and new centroids $\tilde{\underline{\mu}}_1$, and $\tilde{\underline{\mu}}_2$, having special forms

(B.2) $\quad \tilde{\Sigma} = I$

and

(B.3) $\quad (\tilde{\underline{\mu}}_1 - \tilde{\underline{\mu}}_2)' = (\delta_p, 0, \ldots, 0)$ .

The following derivation of the estimator for $E(D_c^2)$ uses the convenience of working with two multivariate normal subpopulations having parameters $\tilde{\Sigma}$, $\tilde{\underline{\mu}}_1$, and $\tilde{\underline{\mu}}_2$. This derivation, however, is not specific to this special type of population. It's applicability to any arbitrary pair of multivariate normal subpopulations is guaranteed by the invariance of $\delta_p^2$ to transformation and translation.

The general expression for $D_c^2$ in (23) reduces to

$$(B.4) \qquad D_c^2 = \frac{\delta_p^2 1_{s1}^2}{\underline{1}_s' \underline{1}_s}$$

for subpopulations having the parameters $\tilde{\Sigma}$, $\underline{\tilde{\mu}}_1$, and $\underline{\tilde{\mu}}_2$, in (B.2) and (B.3). The term $1_{s1}$, is the first element of $\underline{1}_s$ and $1_{s1}^2$ is the first element in the sum

$$(B.5) \qquad \underline{1}_s' \underline{1}_s = 1_{s1}^2 + 1_{s2}^2 + \ldots + 1_{si}^2 + \ldots + 1_{sp}^2 \quad .$$

Over random samples of size N, with group sample sizes $N_1$ and $N_2$, the expected value of $D_c^2$ is

$$(B.6) \qquad E(D_c^2) = \varepsilon \left[ \frac{\delta_p^2 1_{s1}^2}{\underline{1}_s' \underline{1}_s} \right] \quad ,$$

which can be estimated by

$$(B.7) \qquad \text{Est}(E(D_c^2)) = \frac{\delta_p^2 \varepsilon(1_{s1}^2)}{\varepsilon(\underline{1}_s' \underline{1}_s)} \quad .$$

In order to proceed further, expressions for the quantities $\varepsilon(1_{s1}^2)$ and $\varepsilon(\underline{1}_s' \underline{1}_s)$ are needed. Fortunately, Kshirsagar (1972) has derived an estimator for $(N-2)^{-2}\varepsilon(\underline{1}_s \underline{1}_s')$,

$$(B.8) \qquad (N-2)^{-2}\varepsilon(\underline{1}_s \underline{1}_s') = \left[ [(N-3) + \frac{N_1 N_2 \delta_p^2}{N}]\Sigma^{-1} + (N-r-3)\frac{N_1 N_2}{N}\Sigma^{-1}(\underline{\mu}_1 - \underline{\mu}_2)(\underline{\mu}_1 - \underline{\mu}_2)'\Sigma^{-1} \right] G^{-1}$$

where

$$(B.9) \qquad G = N_1 N_2 (N-r-2)(N-r-3)(N-r-4)/N \quad .$$

For subpopulations having parameters $\tilde{\Sigma}$, $\underline{\mu}_1$, and $\underline{\mu}_2$, a simplified expression for $\varepsilon(\underline{1}_s \underline{1}_s')$ can be otained from (B.8),

$$(B.10) \qquad \varepsilon(\underline{1}_s \underline{1}_s') = \frac{(N-2)^2}{G} \left[ [(N-3) + N_1 N_2 \delta_p^2 N^{-1}]I + N_1 N_2 N^{-1}(N-r-3)\Delta_p \right] \quad ,$$

where $\Delta_p$ is a r-by-r singular matrix containing $\delta_p^2$ as its first element and zeros elsewhere.

The relationship (Kshirsagar, 1972)

(B.11) $\quad \varepsilon(\underline{l}_s'\underline{l}_s) = \text{trace}\,[\varepsilon(\underline{l}_s\underline{l}_s')]$

can be used to obtain

(B.12) $$\varepsilon(\underline{l}_{s1}^2) = \frac{(N-2)^2}{G}\left[(N-3) + N_1N_2N^{-1}(N-r-2)\delta_p^2\right]$$

and

(B.13) $$\varepsilon(\underline{l}_s'\underline{l}_s) = \frac{(N-2)^2}{G}\left[(N-3)[r + N_1N_2N^{-1}\delta_p^2]\right] \quad .$$

Substituting (B.12) and (B.13) into (B.7) yields

(B.14) $$\text{Est}(\varepsilon(D_i^2)) = \frac{\delta_p^2[N-3 + N_1N_2N^{-1}(N-r-2)\delta_p^2]}{(N-3)[r + N_1N_2N^{-1}\delta_p^2]} \quad ,$$

an estimate for $\varepsilon(D_c^2)$ in terms of $\delta_p^2$, the population generalized distance. (Note that when r = 1, $\text{Est}(\varepsilon(D_c^2))$ equals $\delta_p^2$. This result is not surprising. When the original subpopulations are unidimensional, a sample discriminant function merely rescales the original dimension and the standardized distance between population centroids along that dimension remains invariant.)

An Estimator for $E(D_c^2)$ in Terms of Sample Quantities

For (B.14) to be usable an estimator for the term $\delta_p^2$ is needed. Either $(N-r-3)[D_s^2 - (N-2)Nr/(N_1N_2(N-r-3))]/(N-2)$ or $[D_s^2(N-r-3)/(N-2)]$ can be used. The former is an unbiased estimator, but Lachenbruch and Mickey

(1968) used the latter in the DS-method to avoid negative estimates of the population generalized distance. Once either term is used in (B.14), the resulting estimate of $\varepsilon(D_c^2)$ can be substituted into (B.1) for $D_c^2$ to estimate $\min(E_c)$, which can serve as a lower bound estimate of $E_c$. The mnemonics given this procedure of estimating through an estimate of $\min(E_c)$ are the MS-method when the biased estimate of $\delta_p^2$ is used, and the MU-method when the unbiased estimate of $\delta_p^2$ is used.

# APPENDIX C

## PROOF THAT MIN($E_c$) IS A FUNCTION OF $D_c^2$

The claim has been made that the minimum actual error rate associated with use of a sample classification rule in the population is a function of the "shrunken" Mahalanobis distance between subpopulation centroids along the dimension defined by the sample discriminant function. This claim is proved here for the case where costs of misclassification are assumed to be equal.

### The Marginal Distribution Along $\underline{l}_s$

Recall that the two subpopulations $K_1$ and $K_2$ follow multivariate normal density functions with centroids $\underline{\mu}_1$ and $\underline{\mu}_2$ and a common covariance matrix $\Sigma$. It is well known that a nonsingular transformation of a multivariate normal population will produce transformed variates which also follow a multivariate normal distribution. Another property of multivariate normal populations is that the marginal density functions are univariate normal with means and variances obtained by taking the appropriate components of $\underline{\mu}_k$ and $\Sigma$ (Anderson, 1958).

The sample discriminant weights $\underline{l}_s'$ can be viewed as one row of a nonsingular r-by-r transformation matrix which reorients the reference frame in the r-dimensional space. The remaining (r-1) rows of the transformation are chosen such that the dimensions they produce are mutually orthogonal and orthogonal to the dimension formed by the discriminant weights $\underline{l}_s'$. (These r new dimensions are statistically independent by virtue of their

normality.) When the two subpopulation centroids $\underline{\mu}_1$ and $\underline{\mu}_2$ are projected onto the dimension defined by $\underline{l}_s'$ and accompanied by the translation, $-\underline{l}_s'(\bar{\underline{x}}_1 + \bar{\underline{x}}_2)/2$, the means

(C.1) $\mu_1^* = \underline{l}_s' [\underline{\mu}_1 - .5(\bar{\underline{x}}_1 + \bar{\underline{x}}_2)]$

and

(C.2) $\mu_2^* = \underline{l}_s' [\underline{\mu}_2 - .5(\bar{\underline{x}}_1 + \bar{\underline{x}}_2)]$

are the end result. Within each subpopulation, the variance along this dimension is

(C.3) $V_w = \underline{l}_s' \Sigma \underline{l}_s$ .

The scores within each subpopulation along this dimension are distributed normally with mean $\mu_k^*$ and variance $V_w$. In formal notation, we say that for each subpopulation k the density function for x* is

(C.4) $$f_k(x^*) = (2\pi V_w)^{-.5} \exp\left[-.5\left(\frac{x^* - \mu_k^*}{V_w^{.5}}\right)^{.5}\right] .$$

Minimum Actual Error Rate Associated with a Sample Classification Rule

The minimum actual error rate associated with a sample classification rule is equal to the minimum total probability of misclassification along the dimension defined by the sample discriminant weights $\underline{l}_s$. To obtain the minimum total probability of misclassification along this dimension, an optimal classification rule along this dimension is needed, that is, the cutoff score must be adjusted. Welch's (1939) solution is as applicable to classification along this single dimension defined by $\underline{l}_s$ as it is to classification in the original r-dimensional space.

General Solution

Let $f_k(x)$ be the density function of x if it comes from subpopulation K. Let $q_k$ be the proportion of the total population that is in subpopulation K. Assign x to $K_1$ if x is in some region $R_1$ and to $K_2$ if x is in a region $R_2$. Assume that $R_1$ and $R_2$ are mutually exclusive and that their union includes the entire space R. The total probability of misclassification is

$$(C.5) \quad E_p = q_1 \int_{R_2} f_1(x)d\underline{x} + q_2 \int_{R_1} f_2(x)d\underline{x}$$
$$= q_1[\int f_1(x)d\underline{x} - \int_{R_1} f_1(x)d\underline{x}] + q_2 \int_{R_1} f_2(x)d\underline{x}$$
$$= q_1 \int f_1(x)d\underline{x} + \int_{R_1} [q_2 f_2(x) - q_1 f_1(x)]d\underline{x} \quad .$$

In order to minimize $E_p$, $R_1$ should be chosen such that

$$(C.6) \quad q_2 f_2(x) - q_1 f_1(x) < 0 \quad .$$

Thus the classification rule is to assign x to $K_1$ if $f_1(x)/f_2(x) > q_2/q_1$ and to $K_2$ if $f_1(x)/f_2(x) < q_2/q_1$. If $f_1(x)/f_2(x) = q_2/q_1$ it is a tossup.

Optimal Classification Along the Dimension Defined by $\underline{l}_s$

The density functions for scores on the dimension defined by the sample discriminant weights are defined in (C.4). Thus the ratio of $f_1(x^*)$ to $f_2(x^*)$ is

$$(C.7) \quad \frac{f_1(x^*)}{f_2(x^*)} = \frac{(2\pi V_w)^{-.5} \exp[-.5(x^* - \mu_1^*)^2/V_w]}{(2\pi V_w)^{-.5} \exp[-.5(x^* - \mu_2^*)^2/V_w]}$$
$$= \exp[-5(x^{*2} - 2x^*\mu_1^* + \mu_1^{*2})/V_w + .5(x^{*2} - 2x^*\mu_2^* + \mu_2^{*2})/V_w]$$
$$= \exp[x^*(\mu_1^* - \mu_2^*)/V_w - .5(\mu_1^* - \mu_2^*)(\mu_1^* + \mu_2^*)/V_w]$$
$$= \exp[(x^* - .5(\mu_1^* + \mu_2^*))(\mu_1^* - \mu_2^*)/V_w] \quad .$$

Taking the natural logarithm yields the optimal classification rule, which is to assign $x^*$ to $K_1$ if

(C.8) $$W_p(x^*) = \left[ [x^* - .5(\mu_1^* + \mu_2^*)](\mu_1^* - \mu_2^*)/V_w \right] > \ln(q_2/q_1)$$

and to $K_2$ if

(C.9) $$W_p(x^*) < \ln(q_2/q_1) \quad .$$

Note that $W_p(x^*)$ has the standard form of an optimal classification rule: the scores $x^*$ are multiplied by a scaling factor $(\mu_1^* - \mu_2^*)/V_w$, which is the univariate expression for the coefficients of Fisher's discriminant function and then this score is adjusted by subtracting the additive constant $(\mu_1^* + \mu_2^*)/2$.

Since $W_p(x^*)$ follows a univariate normal distribution, we are able to calculate the optimal error rate (in this space of reduced dimensionality) by using the cumulative normal distribution function $\Phi(z)$. The probability of a member of $K_1$ being misclassified by $W_p(x^*)$ is

(C.10) $$P_1 = \text{Prob}((W_p(x^*) < \ln(q_2/q_1)) \,|K_1]$$

To use the cumulative normal distribution function, scores on $W_p(x^*)$ must be standardized. Let $z_1^*$ equal the scores $W_p(x^*)$ standardized in the metric of $K_1$,

(C.11) $$Z_1^* = (W_p(x^*) - W_p(\mu_1^*))/(V_w^*)^{.5} \quad ,$$

where $V_w^*$ is the variance of the $W_p(x^*)$ scores. The mean $W_p(\mu_1^*)$ equals

(C.12) $$\begin{aligned} W_p(\mu_1^*) &= [\mu_1^* - .5(\mu_1^* + \mu_2^*)][(\mu_1^* - \mu_2^*)/V_w] \\ &= .5(\mu_1^* - \mu_2^*)[(\mu_1^* - \mu_2^*)/V_w] \\ &= .5(\mu_1^* - \mu_2^*)^2/V_w \quad . \end{aligned}$$

When the relationships in (C.1) - (C.3) are substituted into (C.12),

$$(C.13) \quad W_p(\mu_1^*) = .5\left[\frac{(\underline{1}_s'(\underline{\mu}_1 - \underline{\mu}_2)\underline{1}_s)^2}{\underline{1}_s'\Sigma\,\underline{1}_s}\right]$$

it reduces to the "shrunken" Mahalanobis distance divided by two. An analogous derivation for $W_p(\mu_2^*)$ yields

$$(C.14) \quad W_p(\mu_2^*) = -.5D_c^2 \quad .$$

The variance of $W_p(x^*)$ in either subpopulation, $K_1$ or $K_2$ can be expressed as

$$(C.15) \quad V_w^* = (\mu_1^* - \mu_2^*)V_w^{-1}V_wV_w^{-1}(\mu_1^* - \mu_2^*)$$
$$= (\mu_1^* - \mu_2^*)^2/V_w$$
$$= D_c^2 \quad .$$

In terms of the standardized variable $z_1^*$, $P_1$ can be expressed as

$$(C.16) \quad P_1 = \text{Prob}\left[\left(z_1^* < \frac{\ln(q_2/q_1) - .5D_c^2}{D_c}\right) \,\middle|\, K_1\right] \quad ,$$

which can be rewritten as

$$(C.17) \quad P_1 = \Phi\left[\frac{\ln(q_2/q_1) - .5D_c^2}{D_c}\right] \quad ,$$

where $\Phi(z_0)$ is the cumulative distribution function of normal variable with mean zero and variance unity evaluated at $z_0$, i.e.,

$$(C.18) \quad \Phi(z_0) = \int_{-\infty}^{z_0}(2\pi)^{-5}\exp[-5(z^2)]dz \quad .$$

To determine the probability of misclassifying a member of $K_2$, an analogous derivation is followed, yielding

$$(C.19) \quad P_2 = \Phi[-(\ln(q_2/q_1) + .5D_c^2)/D_c]$$
$$= 1 - \Phi[(\ln(q_2/q_1) + .5D_c^2)/D_c] \quad .$$

The total probability of misclassification can then be expressed

$$(C.20) \quad E_p^* = q_1 P_1 + q_2 P_2$$

$$= q_1 \Phi \left[ \frac{\ln(q_2/q_1) - .5D_c^2}{D_c} \right] + q_2 \Phi \left[ \frac{-\ln(q_2/q_1) + .5D_c^2}{D_c} \right]$$

$$= \min(E_c) \quad .$$

It has just been demonstrated that the minimum actual error rate associated with use of the sample classification rule in the population is a function of the "shrunken" Mahalanobis distance between subpopulation centroids projected along this dimension.

Relationship Between $W_s(\underline{x})$ and $W_p(x^*)$

The rule $W_p(x^*)$ is the optimal classification rule along the dimension defined by the sample discriminant weights. It is interesting to compare $W_p(x^*)$ to the sample classification rule $W_s(\underline{x})$. To make this comparison possible, it is desirable to express $W_p(x^*)$ in terms of the original variables $\underline{x}$. Using the relationships in (C.1) - (C.3), (C.8) can be rewritten as

$$(C.21) \quad W_p(x^*) = \underline{l}_s'[\underline{x} - .5(\bar{\underline{x}}_1 + \bar{\underline{x}}_2)][s] \quad ,$$

where s is the scaling constant

$$(C.22) \quad s = \left[ \frac{\underline{l}_s'(\underline{\mu}_1 - \underline{\mu}_2)}{\underline{l}_s' \Sigma \underline{l}_s} \right] = \sigma_y^2 D_c^2 / (\underline{\sigma}_{yx} \underline{l}_s) \quad ,$$

where $\sigma_y^2$ is the variance of the criterion of group membership and $\underline{\sigma}_{yx}$ is the 1-by-r vector of covariances between group membership and the r predictor variables. Let's define $\underline{d}_s$ as the difference in sample centroids

and $\underline{d}_p$ as the difference in population centroids, such that

(C.23) $\underline{d}_s - \underline{d}_p = [(\bar{\underline{x}}_1 - \bar{\underline{x}}_2) - (\underline{\mu}_1 - \underline{\mu}_2)]$ .

This expression allows us to rewrite (C.21) as

(C.24) $W_p(x^*) = [\underline{l}_s'(\underline{x} - .5(\underline{x}_1 - .5(\bar{\underline{x}}_1 - \bar{\underline{x}}_2)) + .5\underline{l}_s'(\underline{d}_s - \underline{d}_p)][s]$

which expresses $W_p(x^*)$ as a function of $W_s(\underline{x})$, the sample classification rule. Note that the two rules differ by a constant $\underline{l}_s'(\underline{d}_s - \underline{d}_p)/2$ and a scaling factor. When the subpopulations $K_1$ and $K_2$ are of equal size, the only important difference between $W_s(\underline{x})$ and $W_p(x^*)$ is the additive constant, which is a function of how well the sample centroid difference, $\underline{x}_1 - \underline{x}_2$, approximates $\underline{\mu}_1 - \underline{\mu}_2$, the population centroid difference. To the extent that this approximation is good, the actual error rate associated with $W_p(\underline{x})$ will approach the minimum actual error rate associated with $W_p(x^*)$. For example, if $\underline{d}_s - \underline{d}_p = 0$, then $W_s(\underline{x})$ is merely a rescaling of $W_p(x^*)$ by a multiplicative constant 1/s, and the standardized version of $W_s(\underline{x})$ is identical to the standardized version of $W_p(x^*)$.

## APPENDIX D

## RELATIONSHIP BETWEEN $R_c^2$ AND $D_c^2$

Consider applying sample discriminant weights $\underline{l}_s$ to the population scores and computing the correlation between the binary criterion variable of group membership and the scores obtained by applying $\underline{l}_s$ to the r predictors. The sample discriminant weights, are

$$\underline{l}_s = C^{-1}(\bar{\underline{x}}_1 - \bar{\underline{x}}_2) \quad , \tag{D.1}$$

where C is the pooled within groups covariance matrix and $\bar{\underline{x}}_1$ and $\bar{\underline{x}}_2$ are the centroids in samples from subpopulations $K_1$ and $K_2$, respectively. The squared correlation (or cross-validity coefficient) between the binary criterion and the predictions based on $\underline{l}_s'$ is

$$R_c^2 = (\underline{l}_s'\underline{\sigma}_{xy})^2/(\sigma_y^2\underline{l}_s'\Sigma_{xx}\underline{l}_s) \quad , \tag{D.2}$$

which, due to the binary nature of the criterion Y, can be rewritten as

$$R_c^2 = \frac{((\bar{\underline{x}}_1-\bar{\underline{x}}_2)'C^{-1}(\underline{\mu}_1-\underline{\mu}_2)q_1q_2)^2}{q_1q_2\,(\bar{\underline{x}}_1-\bar{\underline{x}}_2)'C^{-1}\Sigma_{xx}C^{-1}(\bar{\underline{x}}_1-\bar{\underline{x}}_2)} \tag{D.3}$$

$$= \frac{q_1q_2((\bar{\underline{x}}_1-\bar{\underline{x}}_2)'C^{-1}(\underline{\mu}_1-\underline{\mu}_2))^2}{(\bar{\underline{x}}_1-\bar{\underline{x}}_2)'C^{-1}\Sigma_{xx}C^{-1}(\bar{\underline{x}}_1-\bar{\underline{x}}_2)}.$$

Note from (A.9), that $\Sigma_{xx}$ can be expressed as the sum of a within-groups and a between-groups covariance matrix. Thus, (D.3) can be rewritten as

$$R_c^2 = \frac{q_1q_2((\bar{\underline{x}}_1-\bar{\underline{x}}_2)'C^{-1}(\underline{\mu}_1-\underline{\mu}_2))^2}{(\bar{\underline{x}}_1-\bar{\underline{x}}_2)'C^{-1}[\Sigma + q_1q_2(\underline{\mu}_1-\underline{\mu}_2)((\underline{\mu}_1-\underline{\mu}_2)']C^{-1}(\bar{\underline{x}}_1-\bar{\underline{x}}_2)} \tag{D.4}$$

$$R_c^2 = \frac{q_1 q_2 ((\bar{x}_1-\bar{x}_2)'C^{-1}(\mu_1-\mu_2))^2}{(\bar{x}_1-\bar{x}_2)'C^{-1}\Sigma C^{-1}(\bar{x}_1-\bar{x}_2) + (\bar{x}_1-\bar{x}_2)'C^{-1}q_1q_2(\mu_1-\mu_2)(\mu_1-\mu_2)'C^{-1}(\bar{x}_1-\bar{x}_2)} .$$

Note the two scalar quantities

(D.5) $$(\bar{x}_1-\bar{x}_2)'C_{xx}^{-1}(\mu_1-\mu_2) = (\mu_1-\mu_2)'C_{xx}^{-1}(\bar{x}_1-\bar{x}_2)$$

are equal to each other and define the difference between centroids projected onto the discriminant dimension defined by $l_s'$. Hence, (D.4) can be rewritten as

(D.6) $$R_c^2 = \frac{q_1 q_2 \,(\bar{x}_1-\bar{x}_2)'C^{-1}(\mu_1-\mu_2)^2}{(\bar{x}_1-\bar{x}_2)'C^{-1}\Sigma C^{-1}(\bar{x}_1-\bar{x}_2) + q_1 q_2 \,(\bar{x}_1-\bar{x}_2)'C^{-1}(\mu_1-\mu_2)^2} .$$

Rearranging terms in (D.6) yields

(D.7) $$R_c^2[(\bar{x}_1-\bar{x}_2)'C^{-1}\Sigma C^{-1}(\bar{x}_1-\bar{x}_2)] = q_1q_2[(\bar{x}_1-\bar{x}_2)'C^{-1}(\mu_1-\mu_2)]^2 \,(1-R_c^2)$$

and

(D.8) $$\frac{R_c^2}{(1-R_c^2)} \Big/ (q_1q_2) = \frac{(\bar{x}_1-\bar{x}_2)'C^{-1}(\mu_1-\mu_2)^2}{(\bar{x}_1-\bar{x}_2)'C^{-1}\Sigma C^{-1}(\bar{x}_1-\bar{x}_2)} .$$

Upon noting the definition of $l_s'$ in (D.1), D.8) can be rewritten as

(D.9) $$\frac{R_c^2}{(1-R_c^2)} \Big/ (q_1q_2) = \frac{l_s'(\mu_1-\mu_2)^2}{l_s'\Sigma l_s} .$$

The expression on the right hand side of the equality is the "shrunken" Mahalanobis distance between population centroids along the dimension defined by $l_s'$. Hence, it has been shown that

(D.10) $$D_c^2 = \frac{R_c^2}{(1-R_c^2)} \Big/ (q_1q_2) .$$